

VISIT...

LANZAROTE
Caliente.COM

4

Information: Harnessing the Data

DATA STORAGE

Having left our flow cytometry system with light signals from a single cell recorded in appropriate channels on the analog-to-digital converter (ADC), we are now faced with the prospect of losing all these data as soon as we start recording data from our next cell. What is required is a way to store the data permanently for later correlation and analysis at our leisure. At this point we leave cytometry *sensu stricto* and find ourselves in the realm of computer ware (soft and hard).

The main challenge encountered with storage of data from flow cytometry arises from the ability of a flow system to generate large amounts of data very quickly. In a four-parameter configuration (four photodetectors), each particle flowing past the light beam generates four signals. If we want to analyze 10,000 cells from each sample (this sounds like a lot to someone used to microscopy, but flow cytometrists can analyze 10,000 cells in, say, 10 seconds, and therefore are easily persuaded that a large number of cells gives statistically better information), then each sample that is run through the flow cytometer will generate 40,000 numbers. If each of those numbers covers the range of 0–1023, then 10 bits ($2^{10} = 1024$) of information are required to specify each of those numbers. Because bits come in packets of 8 (and 8 bits are known as one “byte”), we need two bytes of storage space to specify the intensity of each parameter. This comes to 80,000 bytes for a file that describes four parameters about each of 10,000 cells (plus a few extra bytes for housekeeping

arrangements). A six-parameter cytometer will generate proportionally more data, that is, 120,000 bytes from that same sample.

Therefore (staying with our downmarket four-parameter data file), the information from just seventeen 10,000-cell samples will fill a 1.4 MB floppy disk. Floppy disks are readily available and may be the answer to data storage problems if the experiments are small; they will be an expensive and cumbersome answer if the experiments are large. Although all flow cytometrists start out with small experiments, most progress rapidly to experiments with 20 or 30 or more samples (think of using 96 well microtiter plates for processing cells). In fact, one of the surest rules of flow cytometry (a rule even better known to imaging scientists) is that data will, in less time than predicted, fill all available storage capacity. When you find yourself continually running out of floppy disks, you will begin to try to think of other options for data storage. The other options for flow cytometry data storage are just the same as for any other kind of computer data storage: These options change with time, but currently include, for example, hard drives, zip cartridges, and CD-ROMs (Table 4.1). Options at any given computer will be determined or limited by the peripheral hardware available, but most systems will have a large-

TABLE 4.1. The Number of Flow Data Files (10,000 Cells/4-Parameter Data/1024-Channel Resolution) that Can Be Stored to Various Types of Media^a

Medium	Capacity	Cost (US\$)	Cost per MB (US\$)	Number of files per disk
Floppy disks	1.4 MB	\$0.50	\$0.36	17
Zip cartridges	100 MB	\$10	\$0.10	1,250
	250 MB	\$15	\$0.06	3,125
Jaz cartridges	1 GB	\$100	\$0.10	12,500
Hard drives	10 GB	\$300	\$0.03	125,000
	50 GB	\$1200	\$0.02	625,000
CD-ROMs	600 MB	\$1	\$0.0017	7,500

^aThe "number of files per disk" is given as the number of samples (of 10,000 cells each) whose list mode data (4-parameter/1024-channel resolution) after acquisition can be stored to media of the indicated size. The capacity in bytes of the different media is representative but will vary with the formats of different computing systems. Similarly, different acquisition software will require more or less extra storage space for the housekeeping information that is stored with each sample. Prices of media are illustrative, but will vary considerably from place to place and over time.

capacity hard drive, which might store the information from many samples. Although increasingly inexpensive, hard drives have two main drawbacks. The first is that you cannot take them home with you, and therefore someone else who has access to the system can trash your data (and, given enough time, probably will do just that). The second problem with a hard drive is derived from that rule about data filling all available storage capacity: No matter how large the capacity of the hard drive, it will become full sooner than expected.

The solution to both of these problems is to have removable back-up capability. This back-up device could consist of zip cartridges or floppy disks; both of these options are appropriate for immediate data storage. For long-term archiving, the least expensive (and slowest) option is computer tape. Recordable CD-ROMs currently combine low cost, moderate speed, and a reputation for stability; the price of rewritable CD-ROM burners has come down, but generally this medium has been used for permanent archiving on write-once disks. For example, when six zip cartridges have been filled, their data could be transferred permanently to a CD-ROM and the zip cartridges re-used. Considerations in choice of medium will involve the cost of the medium, the cost of the drive, the speed of writing and accessing the data, and, importantly, the convenience of organizing data files on disks of various sizes. With any luck, you will have backed your data from the hard drive to the removable medium of choice and will have the data safely in your pocket when someone “accidentally” clears the system. For these reasons, back-up capability of some type is especially necessary for multiuser flow cytometers.

DATA ANALYSIS

Now that data have been stored, we come to analysis, which is the real point of everything we have done so far. Methods for data analysis vary. They vary with the inclinations of the software programmer; they also vary with the budget of the cytometer facility. They may be strictly commercial, or they may be homemade. These days, commercial manufacturers of cytometers compete with each other on the basis of their software systems as much as on the basis of their cytometer technology. In addition, independent entrepreneurs, with increasing frequency, have begun to program for analysis of

flow data and have successfully entered into the commercial market. Moreover, there are people (for example, various scientists at the Los Alamos Labs in New Mexico or Joe Trotter at the Scripps Research Institute, San Diego, California) who have developed flow analysis software that they distribute without charge. It is increasingly true that the software available for analysis plays a large role in what the user sees as an acceptable cytometer package. Samples may be run through a cytometer and the information from those particles stored very quickly; analysis and re-analysis of that information may then require a great deal of time. Therefore, it is not surprising that software is an important aspect of flow cytometry.

In theory, data from all flow cytometry systems are now stored in so-called flow cytometry standard (FCS) format. This means that, although data stored after acquisition on one cytometer may not be analyzable on software from another cytometer (because manufacturers have been discouragingly slow at fully embracing the standard), the format is one that can be learned, and, in principle, anyone with good programming skills could write software for analysis of flow cytometry data. The FCS format also means that independent programmers can and do write programs that will handle data acquired on any cytometer. In practice, most people use commercial packages for data acquisition and storage because these packages are commonly purchased along with the cytometer. Although the same packages provide methods for data analysis, there are times when additional analysis software from an independent source may be helpful. This might be to provide advanced analysis methods, to provide especially pretty pictures for slides and publication, to store data to a database, or to allow analysis on one or another home computing system.

The data stored in FCS format are usually “list mode” data. As described above, this means that, in a four-parameter cytometer, four numbers are stored for each cell. A 10,000 cell data file will consist of a long list of 40,000 numbers, with each set of four numbers describing each cell in the sequence in which it passed through the laser beam. By retrieving the stored data, each cell can be analyzed again, and the intensity of each of the four signals for that cell will be known and can be correlated with each other or with the intensity of the four signals from any (or all) other cell(s). This type of list mode

data is useful because no cytometric information has been lost, and it can all be examined again in future computer analyses.

There are other types of data storage that have the advantage of requiring less storage space. So-called single-parameter data storage involves the storage of the intensity profiles for the population of cells in a sample for each parameter separately; the only information stored is, for example, the distribution of forward scatter signal intensities for the cells in the sample; the distribution of side scatter signal intensities for the cells in the sample; the distribution of red fluorescence signal intensities for the cells in the sample; and the distribution of green fluorescence signal intensities for the cells in the sample. In this case, however, no information has been stored about whether the bright green cells are the cells that are bright red or whether they are the cells that are not red. With this kind of storage, we will not know whether the cells with a bright forward scatter signal are red or green or both red and green. Therefore, if we have stored data as single-parameter information, we tie up little storage capacity but we severely restrict our options for future analysis. Unless storage capacity is very limited and the information required from data analysis is also very limited, list mode data storage is by far the best and indeed the only recommended option. Another rule about flow cytometry data analysis is that you are always going to want more information out of a sample than you thought was required when you planned and carried out the experiment. So use list mode data storage unless there is a very good reason for doing otherwise.

Having stored list mode data for all the particles in a sample, software allows the correlation of the data in all possible directions. We can readily look at any one parameter in isolation and analyze the light intensity histogram from the cells in a sample with respect to that parameter. Because we have stored the channel numbers characterizing the signals from each cell, the software can plot a histogram (number of cells at each channel as in the height histogram; see Fig. 3.10) with all the cells placed according to the channel number characterizing the intensity of their signals. In this way we could look at the intensity distribution of, for example, green fluorescence signals from 10,000 cells in one sample. And then we could look at the intensity distribution of red fluorescence signals from the same 10,000

cells. We can, in fact, generate a histogram for each of the parameters measured. Some software will plot data according to channel number; other software will convert the channel data and use a “relative intensity” scale (think of the upper and lower horizontal axes in Fig. 3.9). In the latter case, the software is making assumptions about the accuracy, linearity, and amplification gain on the photodetectors. Once we have plotted the histogram distribution (number of cells on the y-axis at each defined light signal intensity on the x-axis), the software will allow us to analyze this distribution to extract certain kinds of information: for example, what percentage of the cells fall within a specified intensity range, what the most common intensity (channel number) is for the group of cells (the “mode” channel), what the mean intensity channel is for the group of cells, or what the median intensity is for that group (Fig. 4.1).

Just how these values are obtained will vary with the particular software in question. “Cursors” or “markers” can be used to define regions of intensity that may be of interest. For example, we could place a cursor so that it separates the low-intensity range of green fluorescence from the high-intensity range of green fluorescence, and we could then ask how many cells fall within the high-intensity range. A value for the percentage of positively stained cells can be determined by placing a cursor at a position defined by the background fluorescence of unstained cells. By convention at the 1–3% level (that is, at an intensity that clips the bright edge of the unstained cells and makes 1–3% of them “positive”), this kind of cursor is usually the best way to describe a mixed population that consists of both unstained cells and brightly stained cells.

If we are, on the other hand, concerned with the changing fluorescence intensity of a uniform population of cells, we would be better served by using mode or median or mean characteristics of that population (the use of a “percent positive” value is, in fact, highly misleading if we are looking at a population of cells that are uniformly but dimly fluorescent). The mode value, being simply the channel number describing the intensity of the most frequent group of cells (the peak of the histogram), may vary erratically and be poorly reproducible if the population distribution is very broad. The mean value will be incorrect if significant numbers of cells are in the highest channel (255 or 1023) or lowest channel (0) of the histogram. The median value for the population is most reproducible because it

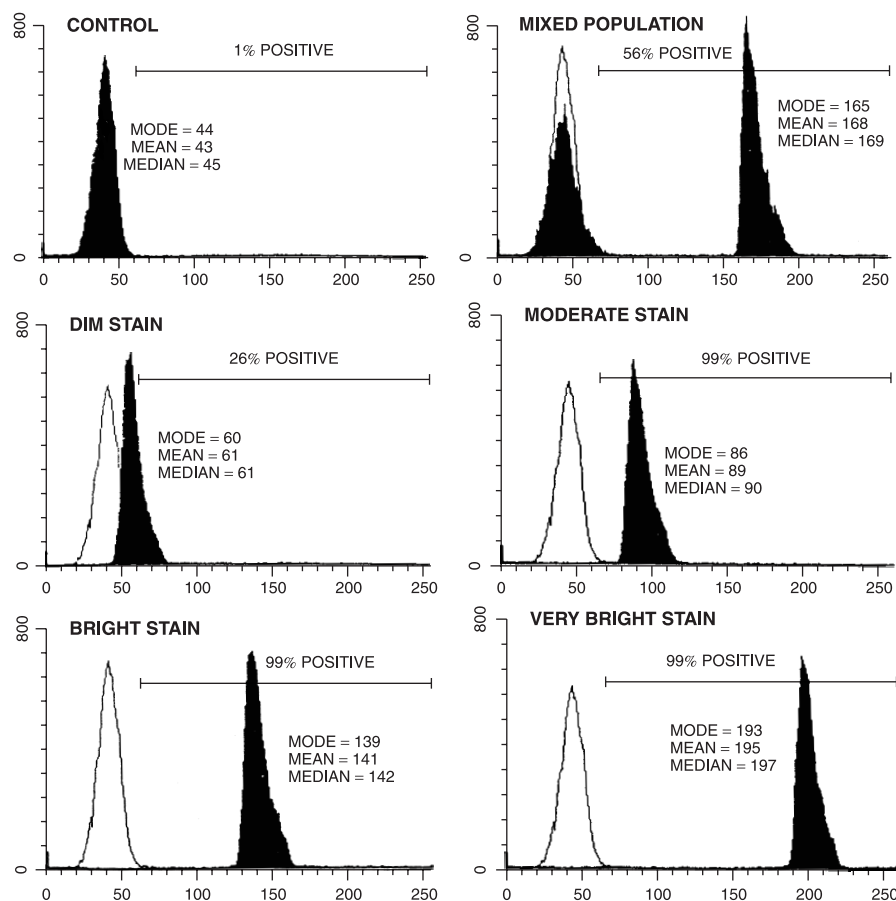


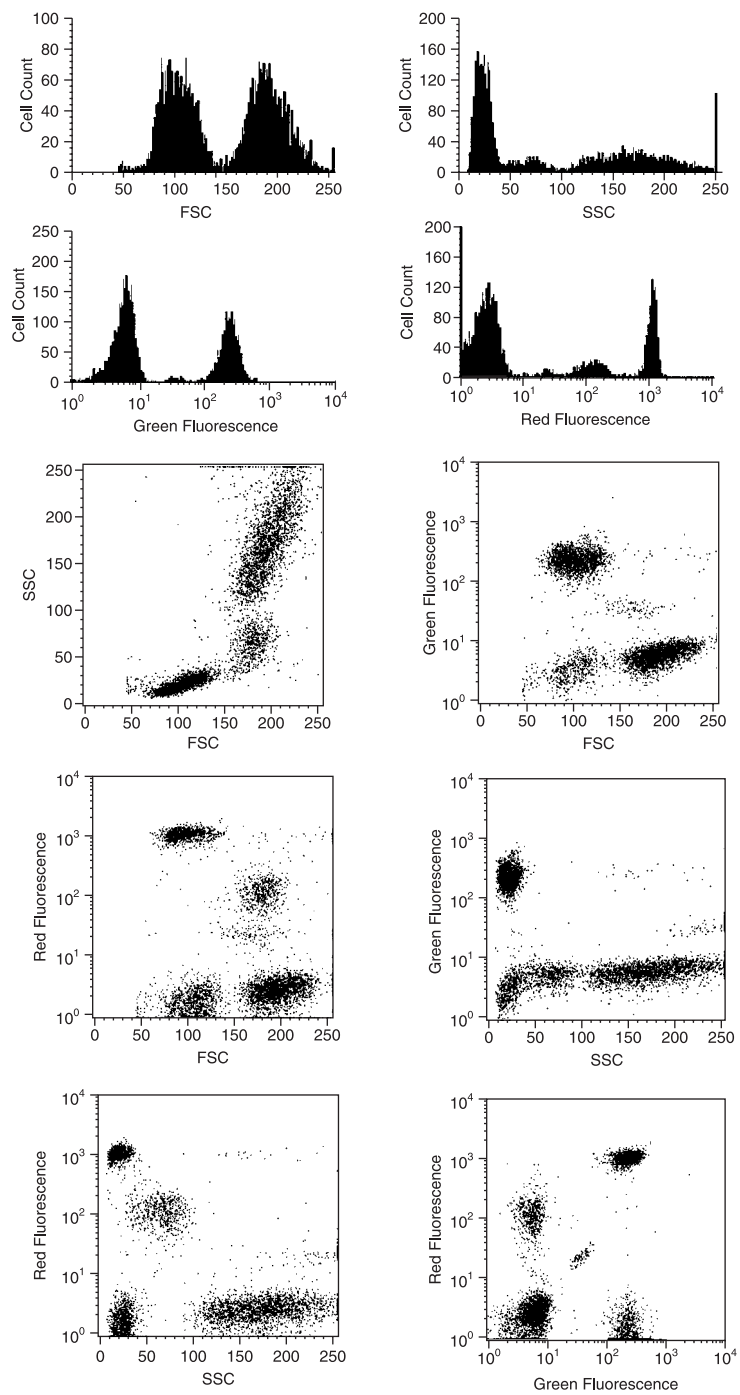
Fig. 4.1. Methods for describing the histogram distribution of signal intensities from a population of cells. The plots show the number of cells on the vertical axis against channel numbers (related to scatter or fluorescence intensity) on the horizontal axis. Control (unstained) cells are indicated as the clear distributions overlaid with the black distributions from the stained cells. If cursors or markers are placed to delineate a region of positive intensity (relative to the 1% level on an unstained control), the “% positive” value can usefully describe a mixed population of stained and unstained particles. This value will be misleading if used to describe a uniform population of dimly stained cells. The “mode,” “mean,” or “median” channel number can be used to compare uniform populations of cells of varying fluorescence intensity.

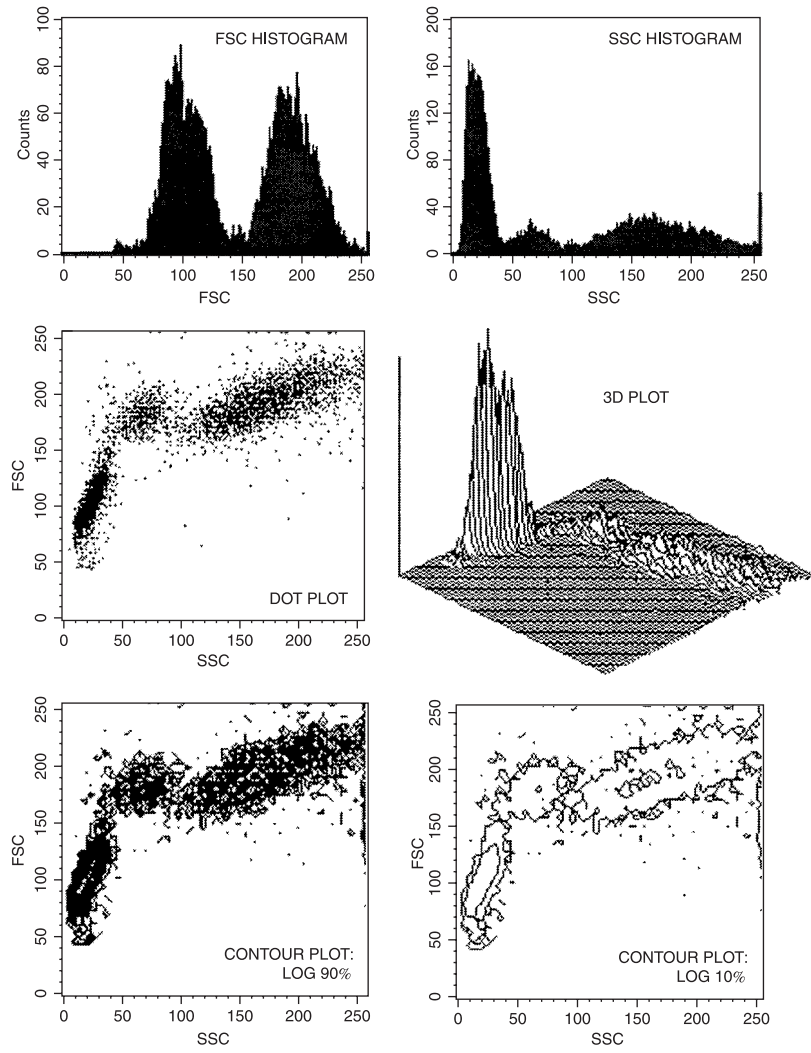
is not affected by small numbers of these “off scale” events or by a few cells with very high or very low fluorescence (“outliers”). For this reason, the median has been recommended by many. However, assuming that there are no events off scale, the mean value will be more directly related to biochemical or bulk measurements of concentration. It should also be added that, if a distribution is symmetrically distributed around the mean, the mean, the mode, and the median will be identical numbers.

If we now want to go further and correlate one parameter with another, software analysis packages implement the drawing of two-dimensional plots. Each cell is placed on the plot according to its intensity channel for each of the selected two parameters. Six two-parameter correlations can be derived from our four-parameter data (Fig. 4.2; keep in mind that each two-parameter correlation could be plotted with the x and y axes reversed). Dot plots show, simply, a dot on the page or screen at each locus defining quantitatively (according to channel number) the two relevant characteristics of each particle in the sample. Dot plots suffer, graphically, from black-out in that an area of a display can get no darker than completely black; if the number of particles at a given point are very dense, their visual impact, in comparison with less dense areas, will decrease as greater numbers of particles are displayed.

Contour plots, as another type of two-dimensional graph, display the same kind of correlation as dot plots, but can provide more visual information about the frequency of particles at any given point in the display. They allow the display of data according to the number of cells in any particular cluster (think of the contour lines describing peaks on a mountaineer’s map). Lines are assigned to various levels of cell density (as contour maps assign lines to different altitudes) according to any one of several different strategies. While changes in the assignment of lines will not change the values calculated for the number of cells in a cluster, they may radically change the way the data set appears. Figures 4.3 and 4.4 show examples of how the same data plotted with different cell density assignments for contour lines

Fig. 4.2. The four histograms and six dual-parameter plots derived from the data from a four-parameter cytometer. Each of the six dual-parameter plots could be drawn with the x and y axes reversed.





A.

Fig. 4.3. Different methods of plotting one set of two-dimensional (forward scatter vs. side scatter) data. **A:** Two separate histograms, a dot plot, a three-dimensional plot, and contour plots according to two different plotting algorithms. **B:** Four additional contour plotting algorithms for the same data.

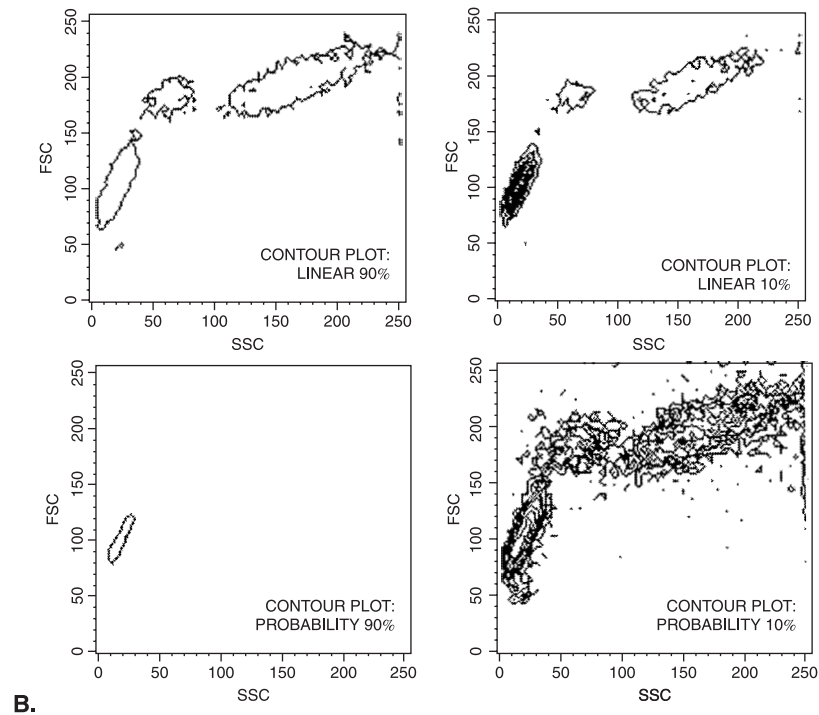


Fig. 4.3 (continued)

can make messy data look tidy (or vice versa) and can even make three peaks look like a single distribution. The message here is simply that we need an informed eye when looking at contour plots generated from other people's data.

Once the data from a sample have been plotted in two dimensions, the distribution of particles can then be analyzed. *Quadrants* dividing the plot into four rectangles is the term used for the cursors applied in two-dimensional analysis. As with one-dimensional histograms, these cursors simply divide the light intensity channels into areas of interest. The number of particles in each of the defined areas can then be counted. A standard analysis procedure might use unstained control cells to define the channels delimiting background red and green fluorescence. Quadrants can then be drawn based on these channels,

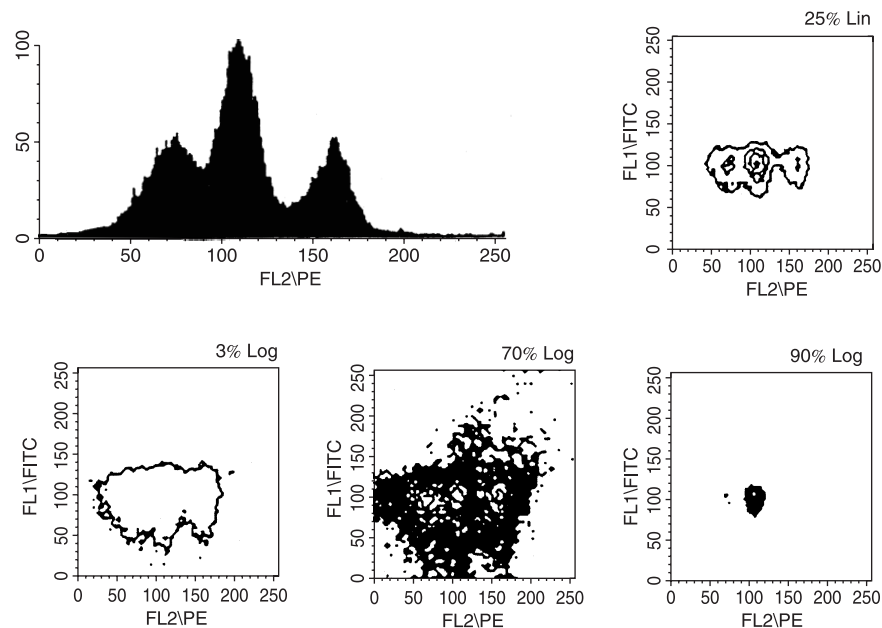


Fig. 4.4. A histogram of fluorescence and contour plots (plotted according to different line assignments) of the same data. Comparison between the histogram and the contour plots allows us to see at what “altitudes” the contour lines have been drawn for each contour plot and why the resulting displays look so different.

and the quadrants will therefore define the staining intensities that we would consider to represent green particles, red particles, and so-called double-positive particles that are both red and green, as well as the double-negative (unstained) cells (Fig. 4.5).

As an intellectual exercise, you should always attempt to visualize the single-dimensional histograms that would result from dot plot or contour plot data. Figure 4.6 indicates dot plots from two different sets of data (with their respective histograms); it is clear that the histograms from these two data sets are identical but that the dot plots are quite different. This example should serve to emphasize that more information is obtained from dual-stained preparations with their two-dimensional plots than from two successive single-dimensional (one color) analyses.

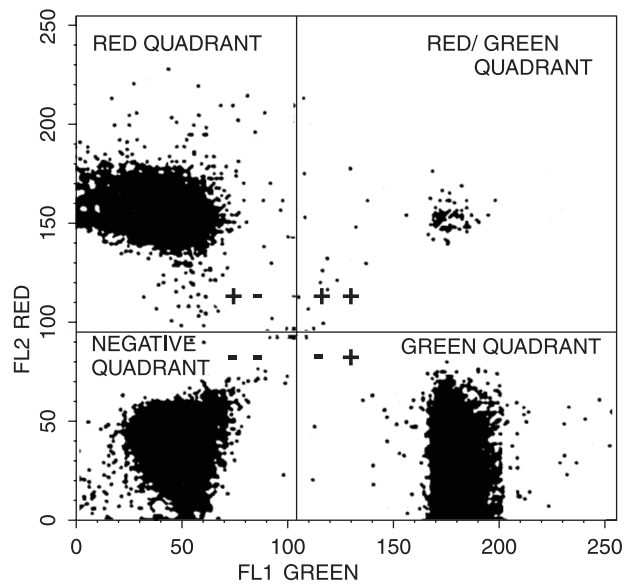


Fig. 4.5. Quadrants are cursors or markers for delineating the intensity (according to channel number) of cells of interest in dual-parameter analysis.

We now need to define *gating*, one of the most subjective aspects of flow cytometry. Gating is based on the defining of *regions*. A region is a way of specifying the characteristics of a subset of particles (in terms of forward scatter, side scatter, and/or fluorescence intensity channel numbers). A region may encompass, for example, all the cells that fall between a certain range of green intensities, or all cells with a certain set of forward and side scatter characteristics, or all cells that are both orange and green. A gate, in contrast, is a combination of regions that define all the cells that we want to include in our final analysis. Cells that pass through the gate get analyzed. A gate can be identical to a single region; that is, the cells we want to include in our final analysis may simply be all the cells that fall into a single, defined region. This is why people often confuse the terms “gate” and “region.” However, a gate can also be defined as a combination of two or more individual regions. For example, a gate may include all the cells that fall into region 1 AND into region 2; or

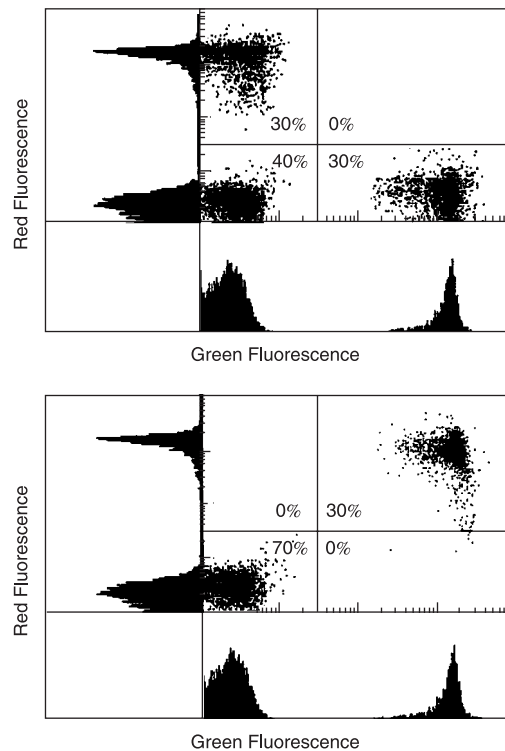


Fig. 4.6. Identical single-color histograms derived from quite different cell samples. The dual-parameter dot plots allow us to distinguish these two populations of cells.

all the cells that fall into region 1 OR into region 2; or all the cells that fall into region 1 but NOT into region 2. If you are young enough, you will have used Boolean algebra in school to combine sets, and these definitions will not present you with any problems. Whether a gate is co-equal to a single region or involves a combination of several regions, any particle that fulfills the defined gate characteristics will pass through the gate and will be included in the next analysis step (Fig. 4.7).

Our gate could, in fact, be a “live gate” or an “analysis gate.” A live gate defines the characteristics of cells that need to be fulfilled before the data from those cells are accepted for storage in a computer for further analysis; the information about all other cells will be lost. For this reason, a live gate should be avoided unless data storage

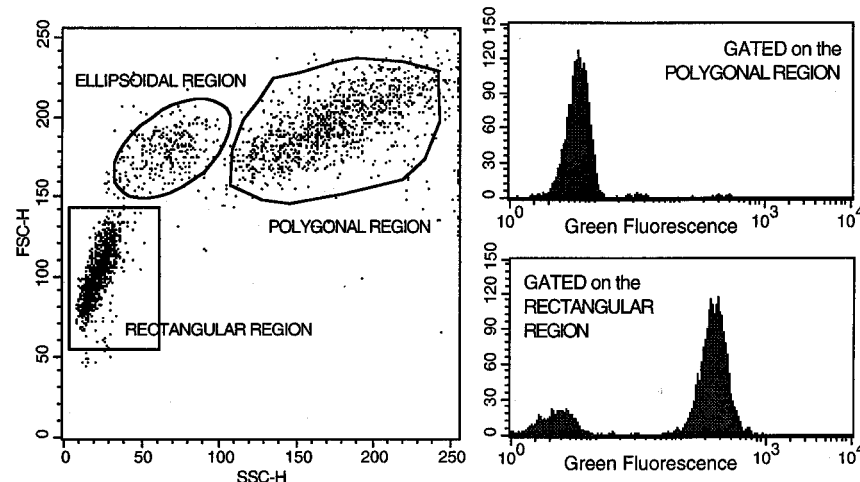


Fig. 4.7. Regions can be drawn to define clusters of cells. Regions can then be used (individually or in Boolean combination) to form gates for restricting subsequent analysis to certain groups of cells.

space is limited. An analysis gate, in contrast, selects cells with defined characteristics from within a large heterogeneous sample of cells, all of whose characteristics have already been stored in a data file. We still have the ability to move the regions around to adapt to different analytical strategies, and we can also calculate the proportion that the gated cells are of the total. The use of either type of gate is rather analogous to the process of, by eye, disregarding all erythrocytes, monocytes, and neutrophils in a microscopic field and then counting only lymphocytes to determine the percentage of lymphocytes that are stained. A trained microscopist's eye is, in that way, defining a lymphocyte gate based on nuclear pattern, shape, and size. A cytometrist can define a lymphocyte's gate in terms of forward and side scatter intensities.

Dual-parameter correlations constitute the standard procedure for analysis of much flow cytometric data. Most cytometers provide us with four or five or more parameters of information. Because humans, in general, are used to thinking in two dimensions, actually correlating three or more parameters with each other can seem rather difficult—in terms of both computer software and our ability to keep track of the strategy (Fig. 4.8). For purposes of data analysis, the most com-

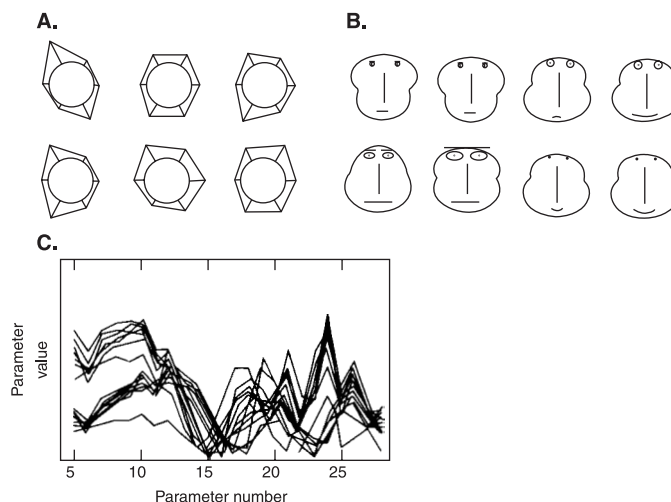


Fig. 4.8. Three different ways to visualize multivariate data. From Dean (1990).

mon procedures use the extra parameters to allow gating (including or excluding) of particles before the final analysis, which is almost invariably just one or two dimensional.

Thus far, we have followed cells into the center of a sheath stream flowing from a nozzle or through a chamber past a focused light beam; we have seen how scatter and fluorescence signals can be generated from those cells when they are hit by that light beam; we have accounted for the registering of those light signals by photodetectors so that the intensities of the signals can be represented by the discrete channels (256 or 1024) of an ADC; we have described the way amplification applied to the output of a photodetector will determine the intensity range represented by the ADC channels; we have seen the way in which the channel number for each of the several parameters characterizing each cell can be stored as list mode data for each cell in sequence in the sample; and we have described the general ways in which the stored data can be displayed and analyzed. We should, therefore, have a clear picture of the fluidic, optical, electronic, and computational characteristics that form the basis for flow cytometric analysis. The chapters that follow will deal with the ways in which these principles are applied to experimental situations. Because the first decisions that are made in designing a flow experiment often

concern the specification of reagents for staining cells, we will initially take a short detour to the principles of photochemistry and laser optics in order to understand the requirements that a flow cytometer imposes on these decisions.

FURTHER READING

Chapter 5 in **Shapiro**, Chapter 30 in Volume I of **Weir**, Section 10 in **Current Protocols in Cytometry**, Chapter 7 in **Darzynkiewicz**, and Chapter 22 in **Melamed et al.** are all good discussions of various aspects of flow data analysis. In addition, an entire book by **Watson (1992)** is devoted to this subject.